DOCUMENT RESUME

ED 464 926 TM 033 856

AUTHOR Miles, Carol A.; Lee, Curtis
TITLE In Search of Soundness in Teacher Testing: Beyond Political Validity.
PUB DATE 2002-04-00
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Foreign Countries; *Political Influences; *Psychometrics; State Programs; *Teacher Evaluation; *Teachers; *Test Use; Testing Programs; *Validity

ABSTRACT

When the drive to develop and administer tests comes from a public call for accountability in a profession, from the government's response to this call, or from politicians attempting to identify a hot issue, it may be said that the test must first have political validity in order to be implemented. Only after this political validity has been achieved, will the real psychometric properties of the test be valued. In addition to introducing the concept of political validity, this paper addresses challenges presented by the traditional psychometric validity concerns surrounding paper-and-pencil tests aimed to measure teacher competencies. This is followed by a review of types of performance-based measures that have been introduced to address validity problems and the psychometric considerations inherent in these tests. Among the tests discussed are the teacher testing program in Massachusetts and teacher assessment as implemented in Ontario, Canada. The review of these and other teacher testing programs suggests that current standardized teacher tests leave much to be desired when it comes to evidence of psychometric validity. The paper concludes that, in spite of psychometric validity, political validity is what is primary in determining a test's public acceptance. (Contains 69 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 464 926

# In Search of Soundness in Teacher Testing: Beyond Political Validity

Carol A. Miles and Curtis Lee

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

C. Miles

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM033856

This paper is prepared for the:
Annual Meeting of the American Educational Research Association in New Orleans, LA
April 2002



BEST COPY AVAILABLE

***In Search of Soundness in Teacher Testing: Beyond Political Validity***

Carol A. Miles, Ph.D. and Curtis Lee, Ph.D.
Faculty of Education, University of Ottawa

Paper Presented at
The American Educational Research Association Annual Meeting
New Orleans, Louisiana, April 2002

## Political Validity

When the drive to develop and administer tests comes from a public call for accountability in a profession, from the government's response to this public call, or from politicians' attempting to identify a "hot issue" which will garner public support, it may be said that the test must first have "political validity" in order to be implemented. Only *after* this political validity has been achieved, (i.e. by the government's and public's acceptance that this test will provide positive public benefits) will the real psychometric properties of the test be valued.

This conception of "political validity" is exemplified by the recent public agenda in the United States, and even more recently in Canada, to improve the outcomes of the public education system through mass testing programs aimed first at students, and then at pre-service and continuing-service teachers (Flippo & Riccards, 2000). Melnick & Pullin (2000) pointed out that the public's (and hence, politicians') concerns demand simplicity, efficiency, and cost-effectiveness of programs. These qualities may be said to run counter to a professional agenda, which would seek more "time-consuming and labor-intensive efforts to implement instructional methodologies, teacher preparation programs, and teacher assessment systems" (p. 265).

Many American jurisdictions have been subjected to this very situation, where politicians realized that the public would be firmly behind any drive to improve public education, and to do this by holding the teaching profession accountable. Recently, the state of Massachusetts implemented a massive teacher testing program, driven initially, according to many accounts, by these political motivations. The outcomes and implications of this program will be further discussed in this paper, and compared with the current state of teacher testing in the province of Ontario, Canada.

In addition to the introduction of the concept of "political validity", this paper will address challenges presented by the traditional psychometric validity concerns surrounding paper and pencil tests aimed to measure teacher competencies will be discussed. This will be followed by a review of the types of performance-based measures which have been developed in order to address these validity problems and the psychometric considerations inherent in these tests. Discussion will then

suggest that regardless of the actual psychometric validity issues of a testing instrument, considerations of political validity will be primary in determining a test's public acceptance.

**The Ontario Teacher Qualifying Test**

In April, 1999, just prior to the launch of a provincial election campaign, Ontario's Premier Mike Harris announced the immediate development and implementation of a province-wide teacher testing program which would use paper-and-pencil tests to determine whether pre-service teachers would receive certification, as well as to make re-certification decisions on practicing teachers on a regular basis (every three to five years). This announcement came as a complete surprise to both the newly established Ontario College of Teachers, and to the Provincial Ministry of Education, who were not consulted prior to the announcement, which met with instant public support, and, not surprisingly, suspicion and outrage from those involved in the teaching profession.

These programs, currently in the development stages, are without precedent in Canada. Critics of these policies (which went from campaign promise to legislation in just over two years) were quick to point out that the Premier's main political advisors had met with the Governor of Massachusetts prior to making the teacher testing announcement, and had identified this as the one issue most certain to garner strong and immediate public support, and it did appear to help in the solid majority won by the Harris government in the subsequent election.

It has recently (February, 2002) also been announced that Ontario teacher education students will be required to write and pass these yet un-standardized tests within weeks of the completion of the tests' development, and that they will not be awarded their teaching certification unless they pass the test. No passing standards have been announced. No study materials are available. To date (5 weeks prior to test administration) neither students nor education faculties have been given any information regarding test content other than the fact that the test will consist of a multiple-choice segment, and a case-analysis segment.

It was not until two months prior to program completion that the students were told that the test would "count"—that they would be required to pass it in order to receive certification to teach. This is leading to extreme test anxiety on the part of the graduating students. Discussion and supposition regarding "The Test", and the anxiety that it is producing has dominated their entire year of post-graduate teacher education. These anxieties would have been considerably alleviated had the Ontario Teacher Qualifying Test been given a standardization year, where students were expected to write it in order to ascertain norming standards only. This was not the case because the Premier, when announcing his original intentions for the test, promised that it would be implemented within a two year period. Therefore, the test was not developed in accordance with accepted test development standards (APA, AERA, NCME, 1999) in order to assure psychometric

soundness and face validity of the test. This could be achieved only by administering it "live" to students after they are made aware of the requirement to pass the test *prior* to entering a teacher education program, and hence making the decision to join the teaching profession..

Instead, in order to fulfill a (very unrealistic) promise to the public, the provincial government is risking the reputation and credibility of their test by administering it in order to make high stakes credentialing decisions before its properties have been psychometrically assessed or shown to be legally defensible.

Despite these apparently damning facts, pre-service teachers, as well as education faculty members find themselves in the untenable situation of not being able to protest this blatant example of the politicization of a test, because the general public holds the belief that if a teacher is qualified, they should be able to pass *the* test—apparently *any* test, that is given to them. Period. *Any complaint on the part of those involved in the education field is seen as exemplary of the defensive and protectionist attitude that, the public would assert, necessitated the test in the first place.* It is a situation for which, it seems, it would be difficult for the government to lose. The test may therefore be argued to have the highest level of political validity.

**Political Validity – An a priori Essential for High Stakes Testing**

Other testing situations may be seen to have similar concerns with what may be labeled "political" validity, especially in areas of high stakes licensure and certification such as the fields of law and medicine, where there are longstanding traditions regarding the types of tests that are administered, and the passing scores which will be accepted and *assumed* by non-experts to protect the public good (Jaeger, 1986). Regardless of the psychometric soundness of these instruments in a changing world of measurement, they must be considered to be generally valid by both members of the professions and the public, a condition which often requires that new (and more psychometrically sound) methods of evaluation be passed up in favour of traditional ones accepted by those who are leaders in their fields (Klass, 1994; Kane, Crooks & Cohen, 1994).

This concern for before-the-fact validity which depends on political motivations, rises out of what appears to be a new role for standardized testing—that of serving to enforce government and public goals for accountability rather than actually assessing specific skills or knowledge of individuals. This is exemplified by the complex political, philosophical, and pedagogical wars currently being waged in the area of teacher testing – specifically for the purposes of this paper, the high-stakes testing of pre-service teachers as a requirement for certification in the profession.

The conception of political validity must not be confused with that of traditional "face validity" (Violato, Marini & McDougall, 1992), which refers to whether test takers, *while writing the test*,

feel that the test is measuring what it is intended for. Instead, political validity refers to whether the general public, or other stakeholders with an interest in the test feel that the test meets their basic requirements and will yield valid results *before* the test is even administered and often without ever seeing the test itself or any evidence of its psychometric soundness. Neither of these two types of validity speak to whether the test scores would yield psychometrically valid results.

Messick (1989) made a major contribution to changing conceptions of validity by the introduction of the concept of consequential validity, which was adopted by the APA/AERA/NCME is its most current Standards (1999). Consequential validity refers to the importance of understanding the impact of test score interpretation. What will the impact of the test score ultimately be, and is this valid based on the test format, content, etc.? Is the use of the test score justifiable based on the properties of the test and testing situation?

As will all currently identified forms of validity, consequential validity is examined AFTER the test has been administered, and scores derived. Political validity may be the only form of validity which is primary prior to test administration. While this seems to contravene currently accepted conceptions of validity as being assessed after-the-fact, it must be recognized that if a test is being administered, even in part, to satisfy someone's demand for a test to be administered *per se*, then the people demanding the test must be satisfied with its qualities prior to its being administered, or the scores will have no *consequence*, as they will not be given credence by the various stakeholders requiring the test in the first place.

Melnick (1996), while not using the term "political validity" described the concept well when suggesting that, despite the best of intentions of the part of legislators when enacting laws which require teacher tests, the singular benefit of the tests themselves may be "a symbolic gesture to assure consumer protection against the absence of absolute minimum competence in beginning teachers." (p. 58). The "symbolism" inherent in this effort should not be underestimated.

From a teacher-testing perspective, this point is especially salient, as it has been argued (Fowler, 2001; Melnick & Pullin, 2000) that the primary motivation for the initial requirement for the test is for politicians to respond to public demands for increased educational quality and accountability. A definitive example of this phenomenon is the current situation in Massachusetts, where the passing standard has been changed several times since the test's implementation in 1998 (when over 60% of candidates failed the initial administration of the test). The state government then acted to threaten that if educational institutions which certified teachers did not produce specific pass rates among their graduates (currently 80%), their programs would be decertified. Teachers' colleges and universities reacted in various ways to this, and consequently, all state colleges now require candidates applying for admission to education programs to take and pass all certification tests *prior* to admission into their programs. All have also begun actively "teaching to the tests". Not

surprisingly, this has led to increases in the pass rates, but is serving to severely limit the demographic of those potential teachers who are applying, and those who are accepted to teacher education programs to a very narrow one (Flippo & Riccards, 2000). The existence of the teacher test in Massachusetts, therefore, has impacted both the admission and curriculum policies of the varies state schools of education, evidencing the political impact that a standardized test can have, irrespective of the test's psychometric properties.

This controversy surrounding the Massachusetts Educator Certification Test (MECT) exemplifies a trend where a test's primary purpose is not to measure *actual* ability on the part of current education graduates but rather to be the impetus for change within the whole teacher education system. (Fowler, 2001). The test's existence is therefore justified to the public by its ability to instigate change within teacher education institutions and the education system itself (Fowler, 2001). Students' individual success or failure on these tests seems to be almost an afterthought in the process.

This is perhaps an example of the most extreme form of political validity, where the actual scores from a test (and the often life-changing pass-fail decisions generated from these scores test's administration) could be seen as secondary to the institutional decisions and policy changes which are generated by the test's *mere existence*. In the case of the Massachusetts test, its primary purpose appears to be that of motivating education programs to assure that they take in candidates with certain qualities, and that they alter their programs to assure that these people pass. If the test's existence assures that these (questionable) actions are taken by the colleges, the test's actual administration (and psychometric soundness, or lack thereof, becomes tertiary (but not tertiary in the lives of those candidates who fail a test and cannot be certified to teach).

The ultimate logic on the part of legislators is that student performance in schools must increase, and that this can be accomplished by flushing out unqualified and illiterate teachers, which, they apparently believe, can be accomplished through relatively simple and inexpensive mass-testing initiatives. In the case of the MECT, this is being achieved through forcing the schools of education to comply by threatening to decertify and institution who don't turn out enough passing graduates. Fowler (2001) refers to this as the "test-and-punish" approach, and contends that it will neither accomplish the goal of increased accountability or better teachers. These actual goals, however, will probably only be measured by a test that is psychometrically sound—both valid and reliable, and the following discussion will attempt to address some of the psychometric issues surrounding various measures of teacher competency.

One quality which leads the teacher tests to be more driven by concerns of political validity than are other professional licensure examinations, however, is that currently in Ontario, as in many American states, the teacher qualifying tests are developed and administered not by the professional

bodies who self-regulate members of their own profession (such as bar associations regulating lawyers, and medical colleges regulating doctors), but rather by the government. Therefore, the content and competencies measured by the test are not those being regulated from within the profession, as is the case with the legal Bar examinations and the Medical Council of Canada licensure examinations, but rather by legislators' and the public's perception of what would adequately measure teaching quality (or, it appears from current rhetoric, how to "put those teachers in their places." This would not be an attitude exemplified by a professional body toward its members. These professional organizations, while having their own previously discussed concerns for the political validity of their licensure tests, also have real concerns for the psychometric validity and reliability of their examinations. This is exemplified by most organizations' employing in-house researchers who primary duty it is to assure the psychometric soundness of their tests.

It must be noted, however, that it doesn't matter how psychometrically valid any professional licensure test is shown to be, if the public and the politicians who represent them don't accept the results as valid, and representing the protection of the public interest, the political agenda will not have been obtained.

**The Psychometric Validity of Teacher Tests – An Historical Perspective**

Despite spending several billion dollars during the 1980s, hundreds of educational jurisdictions in the United States accomplished very little toward the development of psychometrically valid and reliable test instruments of teacher ability. Most of these initial programs, which were put in place in the 1980s and were mostly based on competency/basic skills testing, have subsequently been discarded in favor of newer, more performance based instruments and methods. Some of these programs were designed by major testing firms, and others were custom-developed. Most of these programs did, however, enjoy initial political validity, being both demanded and embraced by the public, and this is what gave the jurisdictions the support necessary to invest so heavily in the tests' development.

An underlying premise which fueled the drive for teacher testing in the United States, and is currently evident in the Canadian effort to develop similar programs, was the public belief that there were large numbers of incompetent, ineffective teachers working in the school system. For the most part, this premise did not arise, either in the United States, or in Canada, from any empirically documented deficiencies in teacher competence, or problems identified by the educational jurisdictions themselves, but from the public's belief that children were being ill-served by the education system, and politicians' eagerness to capitalize on this sentiment.

This is not to say that many of the teacher tests put in place did not serve to raise at least some standards of competency in the profession, but rather that the initial acceptance of them created an

environment where, while not being able to be shown to be psychometrically valid, the tests could be described as "politically valid" because the public believed that the tests would be successful in screening out incompetent teachers. This was, in fact, rarely the case.

Historically, the development of teacher assessment instruments has taken two forms: instruments constructed by professional testing services, which were available to all institutions willing to purchase them, and instruments constructed specifically by some jurisdictions and departments of education for their own use only. These tests were designed to measure four general domains:

***Basic Skills***—ability in reading, writing, and mathematics—have been tested through a large series of standardized tests.
***Subject Area Tests*** attempt to determine the candidates' knowledge in the areas of specialization—or teachable subjects—in most teacher education programs. The assumption behind these tests is that teachers should be able to demonstrate a reasonable level of mastery in subjects that they intend to teach, and that these tests can reliably measure such knowledge (Conklin, 1985).
***Tests of Professional Knowledge and Skills*** attempt to assess knowledge and skills relating directly to teaching which have been gained in teacher education programs.
***Measures of Performance*** are based on actual on-the-job performance, and take the form of everything from simple peer- or supervisor-observation in the classroom to highly structured observations or performance tasks such as computer simulations, in-baskets, and role plays. As the validity of the other tests described above has been brought into question, these assessments are being given more attention.

The primary argument against teacher testing—one that has remained largely unresolved throughout the 20 plus years since its first formal implementation—was that it was difficult, if not impossible, to achieve consensus on a definition of teacher competency. As Piper and Houston (1980) asserted: "Developing valid competency tests is an elusive enterprise. Competence seems so simple when viewed from afar, and so complex when analyzed in detail" (p. 39). Anrig (1987) also argued that because the profession was focusing solely on minimum or basic competencies, the more important, higher order teaching skills were being devalued and ignored. This is an argument that has endured to the present (Darling-Hammond, et al., 2000; Dybdahl, et al., 1997; Klein, 1998).

Even if a universally-accepted definition of teacher competence could be derived, it would have to be empirically proven that the tests administered to teachers were accurate (valid) measures of this construct, which would be, at best, statistically complex, and at worst, impossible, because, as Poggio, et al. (1986) stated:

> Validity refers to the appropriateness, meaningfulness, and usefulness of a certain inference made from a test score (American Psychological Association (APA), 1985). The critical point is that one validates *the use* to which an instrument is put,

> not the measurement instrument. In this light, validity studies must address the particular certification intended, typically as called for in legislation. Therefore, legislation that calls for a test to license "persons expected to be effective teachers" should precipitate studies examining the extent to which test scores differentiate those judged effective from those seen as ineffective (p. 18-19).

The two methods generally used to validate employment tests are known as *criterion related* validation (of the concurrent and predictive types) and *content* validation (Klein, 1998). Criterion-related validity is determined by demonstrating that those who earn high scores on the test are more likely to perform well on the job. This is generally done through assessing correlations with other established tests, or other indicators of job success.

Content validation is generally assured using three basic steps: 1) analyze the job to determine necessary skills and knowledge, 2) assess the relative importance of these skills, and 3) inspect the test to determine whether it measures these skills and knowledge. These determinations are normally made by experts, usually experienced practitioners in the field in question, or those who work closely with them. The major difference between the two types of validity is that criterion-related validity is based on statistical reasoning and can be studied empirically, while content validity relies on subjective judgments (Klein, 1998).

Most of the validity studies relating to teacher testing have been of the criterion-related type. Many deal with comparing candidates' scores on standardized teacher tests to other measures of academic ability.

Many authors still argue that the measurement of professional judgment and pedagogical skill will be the critical factor in identifying teachers who do not meet minimum standards (Dybdahl, et al., 1997). Most teachers who are considered "bad teachers" by the public—those who need to be "weeded out" would not be classified as such by their lack of subject knowledge or basic skills, but rather for poor teaching techniques or human relations skills (Rudner, 1988). These things cannot be validly assessed through simple paper and pencil tests, and the use of paper and pencil tests mayl have the negative impact of giving poor teachers good ratings, and allowing them to challenge principals' poor ratings, and, perhaps, have their shortcomings go undetected, eliminating the possibility of remediation.

The primary validity concerns surrounding the 1980s' paper-and-pencil teacher tests was that they sampled a domain of skills which was not directly related to the craft of teaching, but was more generic in nature. The corresponding, yet perhaps more critical concern was that those abilities and behaviors essential for good teaching were not being measured at all.

Many of these published validity concerns were directed toward the National Teacher Examination (NTE, Educational Testing Service, 1983), which remained the most widely-administered new teacher certification instrument in the United States (Moore, Schurr & Henriksen, 1991) until the introduction of its evolution into the PRAXIS . The NTE was comprised of two independently administered sections—the Core Battery test, which was a basic skills test comprised of three separate tests (communication skills, general, and professional knowledge), and a Specialty Area test which focused on specific content material for the teachable subject, as well as on pedagogical knowledge. There were 42 different specialty areas available, and all tests were in the multiple-choice format.

The Policy Council that developed the NTE claimed that the exam should not be used by either states or schools as a single criterion to assess teacher candidates, as the test was merely a measure of academic skills acquired through formal education, not a measure of how well those skills could be applied in the classroom.

Shepard (1989) warned that this type of paper and pencil test could not capture the complex process of teaching and was limited in what it could measure and what could be concluded from the results and even went so far as to warn that the administration of the test may be counterproductive to the process of classroom teaching, as it may serve to qualify those who were especially good at learning, but were perhaps especially poor at teaching. This warning went largely unheeded by many jurisdictions, however, and as many school districts went about the process of administering the NTE *en mass*, educational measurement researchers were busy conducting an unprecedented number of criterion-related validity studies on the instrument. The reason for this mass administration could be argued to be primarily for the purposes of political validity for a public who could easily identify with, and inherently trusted, the format and were demanding exactly that type of structured testing for teachers.

Written tests would seem to enjoy enhanced political validity because of the perception that they are inexpensive and can be found off-the-shelf and administered to great number of people. Also, they are perceived to provide objective evaluation, something that the public would demand if there was a perception that members of a profession were "protecting their own" It would be assumed that it would be difficult for a professional body to cover for an incompetent teacher or other professional if they received lower than a passing grade on a multiple-choice or short answer test. This, of course, is irrespective of whether the paper-and-pencil test in question is *actually* measuring skills which are required for competence within that profession.

**The NTE and other Competency-Based Tests as General Measures of Academic Ability.** Most of these studies reported that the NTE was little more than a general measure of academic ability and of test taking competency (Madaus & Pullin, 1987; Zimpher, 1990). This was due to

poor predictive or concurrent validity when related to any measure of teacher ability or classroom effectiveness. Many studies did find moderate to strong correlations between the various subtests of the NTE and the prospective teacher's grade point average (GPA, Ayers, 1988; Olstad, Beal & Marrett, 1988; Quirk, Witten & Weinberg, 1973). Moore, Schurr & Henriksen (1991) reported that the scores from the NTE did not improve the ability to predict new teachers' effectiveness as rated by experienced teachers over the use of the GPA alone. This indicated that the use of a prospective teacher's university grades would provide as accurate a measure of teaching abilities as would a score from the NTE, while saving the cost of the administration of the test.

A plethora of studies conducted in the 1980s concurred that scores from these teacher testing instruments correlated highly with, good predictors of GPA, and vice versa (e.g. Ayers, 1988; Dorbry, Murphy & Schmidt, 1985; Egan & Ferre, 1989; Olstad, 1983). Challenges to the psychometric properties of the NTE specifically focused on validity concerns surrounding the lack of demonstrated correlations between either the basic skills competencies or content knowledge and on-the-job performance (Hyman, 1984). The only real predictive abilities of the NTE and other standardized paper-and-pencil tests, then, appeared to be with the candidate's grade point average, and other measures of academic ability, *not* with performance or success as a teacher, or with student achievement (Egan & Ferre, 1989).

As further evidence that the NTE Core Battery was a strong predictor of general academic performance, several authors reported the three subscales of the battery to be highly correlated with each other (Ayers, 1988; Pratt, DeLucia & Williams, 1987; Wakeford, 1988). This would imply that the Communications Skills, General Knowledge, and Professional Knowledge tests are measuring, primarily, the same mental abilities and therefore are not distinguishing between basic skills and teacher-specific skills.

**The NTE as a Predictor of Classroom Performance**. The correlations necessary to confirm that these tests were also good indicators of acceptable levels of teaching ability, however that may be defined, were reported by very few researchers, despite the considerable amount of research directed toward this goal over the past two and a half decades (Dybdahl, et al., 1997).

Stivers and McMorris (1991) raised the issue of the Professional Knowledge portion of the NTE as being harmful to teachers' public reputation as professionals. They argued that, as most of this multiple-choice test appeared to measure mostly general knowledge and common sense, and that if the general public perceived that they could easily pass this test without specialized training, the reputation of teachers as professionals would be jeopardized. This would indicate a dangerously low political validity from the perspective of the teaching profession itself.

Much of the validation research in the area of teacher testing in the United States during the 1980s focused on test bias and the impact of teacher testing on minority teachers. Relatively few studies were directed toward the exploration of the actual ability of these tests to predict teacher competence. Again we see a research focus centered on political concerns, rather than psychometric validity of the teacher tests, although bias toward minority groups would be a psychometric as well as a political concern.

As is indicated by the above discussion, then, the NTE was found to consistently correlate with the new teacher's university grade point average, but not with measures of teacher competence or success. It may therefore be concluded that these paper and pencil tests have not been shown to be valid *predictors* of teachers' performance in the classroom, or of their ability to facilitate student success.

Despite these results, it would be inaccurate to conclude that no useful information can be obtained through the administration of paper-and-pencil type teacher tests. Ayers (1988), for example, reported that the Preprofessional Skills Test (PPST) and the National Teacher Examination (NTE) both offered useful information regarding the basic skills deficiencies of potential pre-service teachers. Similarly, Aksamit and Kluender (1986) found, through national survey techniques, that competency testing was useful because students lacking basic skills were not experiencing success in teacher education programs. This, Aksamit and Kluender argued, led college professors to pay more attention to basic skills in their classes. This again would be an example of how the existence of the test served to drive teachers' college curriculum.

In the wake of the serious validity challenges to the NTE, which were based primarily on its sole reliance on academic and test-taking abilities, and lack of measurement of teachers' unique professional requirements, the PRAXIS Series was developed by Educational Testing Service. This testing series added a third, performance-based dimension to the NTE. The NTE's Core Battery was more or less replicated, in updated format in PRAXIS I: Academic Skills Assessment (also referred to as the Pre-Professional Skills Test, in its paper-and-pencil format). The Subject Area portion of the NTE was reflected in PRAXIS II: Subject Assessments, but with additional subject choices. The new PRAXIS III: Classroom Performance Assessments was developed to address the criterion based validity issues surrounding the previous test.

More recently, some research is indicating that one of the (at least marginally) valid predictors of classroom effectiveness may be encaptured in tests of verbal ability (Fowler, 2001). These contentions have been challenged by Cobb and his colleagues (1999), as they reported that teacher candidates' scores on the Colorado teacher qualifying basic skills reading and writing tests should have correlated strongly with their scores on the SAT and GRE (relationships that have been consistently reported), but did not achieve correlations beyond .5. This indicated a lack of

concurrent validity on the part of the Colorado basic skills tests, and the tests were discontinued. Cobb concluded that those tests were not adequately measuring literacy, and noted that those tests were developed by the same firm who developed and administers the Massachusetts tests—a different firm than that which developed the NTE and the PRAXIS.

While many detractors of teacher testing have reported little evidence that tests of subject-matter expertise are able to identify qualified teachers, Darling-Hammond, Berry, & Thoreson (2001) reported that there was a relationship between student performance in mathematics, and their teachers' holding higher (bachelor's and graduate) mathematics degrees. This relationship was based on the teachers' *attaining* the higher mathematics degrees, however, not on their being given math subject-matter tests prior to certification.

Many authors still argue that the measurement of professional judgment and pedagogical skill will be the critical factor in identifying teachers who do not meet minimum standards (Dybdahl, et al., 1997). Most teachers who are considered "bad teachers" by the public—those who need to be "weeded out"— would not be classified as such by their lack of subject knowledge or basic skills, but rather for poor teaching techniques or human relations skills (Rudner, 1988). These things cannot be validly assessed through simple paper and pencil tests, and the use of paper and pencil tests will have the negative impact of giving poor teachers good ratings, and allowing them to challenge principals' poor ratings, and, perhaps, have their shortcomings go undetected, eliminating the possibility of remediation. This concern leads, then, to the requirement for more complex and comprehensive measures of teaching ability if we are to validly differentiate between effective and ineffective teachers.

## Alternative or Authentic Assessment Instruments

While the multiple-choice style paper-and-pencil tests which make up a significant portion of most teacher qualifying tests have the distinct advantages of being accessible from a variety of test publishers, being relatively inexpensive and easy to obtain, administer, and score, enjoy high objectivity and defensibility of scoring and offer a straightforward comparison of results across test-takers, institutions, and boards, they suffer several disadvantages.

These tests are primarily held to be content valid only, as they have been shown to have little in the way of criterion-related or construct validity, as they are not useful in the prediction of classroom behaviors or teacher success, and have been accused on not assessing what is seen as the "essence" of what makes a good teacher (Ayers, 1988; Dybdahl, Shaw & Edwards, 1997; Garcia & Garcia, 1989; Haney, 1987; Loadman & Brookhart, 1987, 1988; Moore, 1991; Olstad, Beal & Marrett, 1988; Poggio, et al., 1997; Quirk, Witten & Weinberg, 1973; Rudner, 1988). As well, because it is difficult to convince most pre-service or active teachers or administrators that these paper-and-

pencil tests alone can be used to measure all of the abilities and knowledge required for effective teaching, the tests suffer from poor face validity (Shannon & Boll, 1996).

In order to address these shortcomings in the multiple-choice format, several "authentic assessment" techniques—such as structured classroom observations, and various performance-based tasks—have evolved, and are currently being validated. As well, many of the time-tested methods of evaluating teacher performance—such as the principal's evaluation –are still the primary means of assessing ongoing teaching efficacy.

**Direct Observation.** Of these methods, perhaps the least sophisticated and most commonly applied is that of direct observation, where teaching is assessed in the context, which it occurs, and teacher-student interaction can be assessed. With this method, a trained, independent rater visits the classroom to observe the teacher during an actual lesson or lessons. This type of on-the-job evaluation is commonly incorporated into new teachers' evaluations from their mentors or principals, in which case independent raters are not generally used, but, instead, a peer or the principal is the observer. The level of training and expertise of these observers will vary greatly between and within different groups of observers (teachers, principals, professional inspectors, etc.)

One very formalized example of a classroom observation instrument is the System for Teaching and Learning Assessment and Review (STAR, Ellett, Loup, & Chauvin, 1990), which was designed as a comprehensive, classroom-based observation system intended to assess both effective teaching and student learning. This instrument was also designed to assess interactive psychosocial and physical aspects of the overall learning environment (Ellett, Loup & Chauvin, 1991). A highly formalized and extensive training program is available for those who will be assessors under the STAR system.

Validity (predictive and concurrent) studies of the STAR have yielded mixed results, but have generally shown more correlation with student learning and engagement indicators than with student achievement or teacher effectiveness measures. While the instrument did not correlate highly with the results of paper-and-pencil tests, concurrent validity was reported with other classroom observation measures (Ellett, Loup, & Chauvin, 1991). The instrument was shown to have good predictive validity to distinguish between exceptional/ outstanding teachers, as nominated by their peers and supervisors for teaching awards from those receiving no nominations. The ability of the instrument to make the fine distinction between those candidates who would minimally pass and minimally fail on the STAR as compared to other measures has not been established, however.

Matthews, Holmes, Vickers, & Corporaal (1998) addressed concerns regarding the reliability of school inspectors' ratings of classroom behaviors and teaching quality in a study of primary and secondary schools in England. Despite previously poor reliability and validity reports from the

school inspection exercise, they found that two trained inspectors, independently observing the same lesson, identified the same strengths and weaknesses in teaching, and arrived at the same overall conclusions regarding quality. They emphasized the importance of training the inspectors, and the detailed structuring of the observation exercise.

While a few other direct observation instruments such as the North Carolina Teaching Appraisal Instrument (TPAI), are also highly developed like the STAR, most teacher observation instruments are very informal (Gellman, 1993). Based on this general level of informality, Shannon and Boll (1996) pointed out several drawbacks to the observation method. The primary one was the difficulty in arranging observation for teachers prior to their obtaining teaching positions, which may allow poor teachers to be in the classroom for one or two years before being assessed and being given remediation. Another concern related to the problem of one observation's not being sufficient to sample a teacher's full repertoire of skills (and perhaps shortcomings).

These limitations raise the more significant issues of cost and time inherent in having trained raters observe teachers in the classroom. Serious reliability and validity issues have also been raised in regard to these methods. It is of concern that two different raters would not assign the same, or similar, evaluations to the same observation, as well as the concern that teachers will not exhibit regular behaviors when being observed by outsiders. This presence of the rater in the classroom would also, presumably, have an effect on the students' behaviors and teacher/student interaction (Roberson, 1998).

It has long been an axiom of classical test theory that a measure cannot be considered valid if it is not reliable (Crocker & Algina, 1986) . It is in these concerns for the reliability/consistency of ratings that most concerns regarding classroom observation revolve. Roberson (1998) suggested that this poor reliability is primarily due to a lack of formal and rigorous observer training (Roberson, 1998). It is this training activity itself, along with the development of observers' checklists, which can transform classroom observation from an intuitively valid, yet extremely informal task with low reliability into a highly formalized, standardized procedure, which has good reliability. The financial and personnel cost of this transformation, however, can be substantial when comparing the method to other evaluation techniques.

Similar to criticisms of the multiple-choice format, several authors have also criticized the classroom observation method for providing a limited view of teaching, and failing to acknowledge the interrelationships of content-knowledge, pedagogical skill, and situational factors such as school environment (Bird, 1990; Darling-Hammond, 1988, 2000; Scriven, 1988; Shannon & Boll, 1996).

In response to the shortcomings of both of the above-described formats, Stanford University's Teacher Assessment Project (TAP), under the direction of noted educational measurement

researcher Lee Shulman (1989) recommended the use of alternative approaches such as portfolios and simulated performance assessments. Many other educational researchers have argued that a multi-method assessment approach to the evaluation of teaching, which incorporates the use of portfolios and simulations would provide advantages over the traditional methods (Cruickshank & Metcalf, 1993; D'Costa, 1993; Shannon & Boll, 1996). As these methods are discussed below, however, it will become evident that they, too, have certain limitations.

**Portfolios**. Portfolio evaluation requires that teachers gather supporting evidentiary materials from a multitude of sources including, but not limited to lesson plans, teaching materials, journals (professional and personal), supervisor's evaluations, audiotapes, and videotapes. The information provides a cumulative record of a teacher's development and accomplishments over time. Since the early 1990s, in the wake of the falling credibility and application of the paper-and-pencil style of teacher assessments, portfolios have gained increasing acceptance worldwide as a method which is appropriate for the assessment of both preservice and inservice teachers (Barton & Collins, 1993; Bird, 1990; Cole, 1992; Gellman, 1993; Ryan & Kuh, 1993). The National Board for Professional Teaching Standards reported in 1992 that portfolios were also being used extensively for national certification.

Shannon & Boll (1996) discussed several advantages which portfolios hold over the more traditional methods of teacher assessment. They argued that this method has significantly increased face validity over other methods, as they allow the teacher to reflect on the context in which their teaching occurs. Portfolios allow teachers to be involved throughout planning and implementation phases of the assessment itself. The method is applicable to teachers from a wide range of grade levels and subject areas. Teachers can use the portfolios to plan and implement professional development programs. There is also rich opportunity for teachers to interact with each other in the process of developing their portfolios.

The primary disadvantages which have been associated with portfolios are the time it takes to compile and to assess them, the costs involved in this method of evaluation, and the consistency of assessment across different teachers, schools, and boards. The time and labour inherent in the process makes it a costly one not only for the districts using them as evaluation tools, but for the schools whose teachers require time to prepare them. Possible inconsistencies in evaluation are due either to the personal nature of the portfolio, which may be designed to capture very different teaching contexts for different individuals, or to inconsistent ratings by those evaluating the portfolios. These inconsistencies can threaten the reliability of this assessment tool, making it difficult to use the method when comparing individuals or organizations. As well, there are validity concerns, as teachers will normally choose to include in their portfolios only those things, which portray them in a positive light. Shannon and Boll (1996) believed that these concerns would

make it unlikely for portfolio assessments to reach wide acceptance for broadly administered purposed such as certification.

Gellman (1993) expressed the opinion that the benefits of using portfolios could be realized in teacher education programs, and for professional growth regardless of the psychometric soundness of the process. She argued, however, that if portfolios are to be used as an instrument in making certification decisions for new teachers, formal steps must be taken to assure acceptable levels of validity and reliability for the process.

In order to assure these acceptable levels of validity and reliability, it is important to determine what aspects of prospective teachers' performance should be measured, what type of evidence would best represent this proficiency, how the work will be rated, who the raters will be and how they will be trained (Gellman, 1993). These considerations can help to make portfolio assessment a valuable tool for all phases of teacher assessment.

**Teacher Work Samples**. Teacher work samples may be considered a type of limited portfolio evaluation. McConney, Schalock & Schalock (1998) addressed a project in the state of Oregon, which required prospective teachers to design, develop and implement "teacher work samples" as evidence of their professional effectiveness. In order to receive an initial teaching license, these samples must be judged as successful in fostering student learning.

This evaluation required the student teacher to document a three- to five-week period of work in the classroom. The task was grounded in theory of teacher and school effectiveness, competence, and accomplishment, which are both student performance-based and context dependent. McConney, Schalock & Schalock (1998) described the methodology as a complex project, which required careful training of evaluators to retain reliability. The method was considered valid when compared to several other performance based instruments. These authors pointed to the complexity and expense of the process as possible drawbacks, and cautioned that the work samples should comprise part, but not all, of a teacher licensure process.

**Simulation Exercises/Performance Assessment Exercises**. Less expensive than direct observations, these assessments intend to offer a more realistic assessment of teaching than do the pencil-and-paper tests by attempting to simulate everyday tasks that teachers will be required to perform. They include, but are not limited to, in-basket tests, computerized planning simulations, interactive teaching planning and decision-making simulations, micro-teaching, case analysis, role playing, and other group and individual activities. Because teachers perceive the tasks to be relevant, this method has greater face validity than traditional written tests.

The wide variety of possibilities for simulation exercises both enhances and weakens the argument for using this type of assessment (Shannon & Boll, 1996). One difficulty in discussing research into the efficacy of simulations/performance assessments is that many of the different types of simulations can not be compared to each other, and some will be inherently more valid, and have more reliable scoring possibilities than others.

Some examples of performance assessment tasks, which are being integrated with semi-structured interviews (see discussion below) are described by Pecheone and Carey (1990) as being including in Connecticut's Teacher Assessment Center Project (CONNTAC). Some of these tasks include: structuring units presented on index cards, selection of the appropriate approaches to best teach specific concepts and evaluating sample student assignments.

The cost of development of the simulation method of evaluation falls between the less expensive group administered paper-and-pencil tests, and the very expensive individual classroom observations. Simulations require large amounts of time to develop and validate, as well as to administer. Group-administered simulations will tend to be more cost- and time-efficient than those that need to be administered individually. Some simulations will require that personnel be trained to administer them, while other, less costly, ones may be self-administered. The cost, time, and reliability of evaluation will also be dependent on the type of exercise. Some can be designed to be computer-scored, while many others may need to be individually evaluated by expert raters which again implies extensive, and expensive, rater training in order to assure reliability (Shannon & Boll, 1996).

While they appear to be inherently more consequentially valid (Messick, 1989) than paper-and-pencil tests, and more cost-efficient than classroom observations, performance/alternative assessments still confront opposition from within the educational community. As well as the obvious issued of increased cost, these assessments do not produce standardized results, as they are designed to evaluate teachers' competence within the context that the teaching occurs. This makes the outcomes difficult to compare across individuals, schools, or districts, making it difficult to establish concrete pass/fail standards on a wide-scale basis, and making decisions based on these instruments difficult to legally defend.

Shannon and Boll (1996) asserted that while the results of authentic assessment will be meaningful for the teacher, they may be difficult to summarize so that they are easily understood by the public. A cynical public may view these results as an attempt on the part of school administrations to somehow muddy or mask poor basic skills or subject competencies, directly creating a negative impact on the political validity of the instrument (Cruikshank & Metcalf, 1993; Madaus & Kellaghan, 1993; Worthen, 1993).

The cost and time involved in this type of assessment is also of concern to the public. Portfolio assessments and simulations typically cost more to develop than other methods, and questions are raised about whether these costs are justified by the benefits (primarily increased content/consequential validity). Ongoing teacher assessments in England and Scotland which used portfolios and simulations were discontinued due to increased costs, and delays in implementation due to the lengthy nature of test development and validation (Madaus & Kelleghan, 1993).

Pecheone and Carey (1990) summarized the work that the Connecticut Teacher Assessment Center Project had been doing in the development of highly structured teacher performance assessments. They reported that a primary validity concern was that content validity studies should include attention to what *should* be performed on the job, instead of what is usually done. For example, they felt that skill dimensions which may not have traditionally been associated with teachers, such as the ability to explain their pedagogical reasoning may be emphasized, while less emphasis is placed on tasks such as the ability to implement direct instruction, which was not as useful in discriminating between competent and not competent teachers. They warned that a risk in considering this type of "systemic validity" was the concern of fairness when testing for a skill that the majority of practicing teachers may not possess. The underlying intention is to create a test which is fair to individuals, but at the same time encourages the improvement of teachers' abilities and institutional expectations for teachers, again injecting a concern for political validity into the overall validity concerns of the test.

Crocker (1997) raised new validity concerns inherent in the use of performance assessment tasks for certification. The complexity of the exercises, supplemental materials and formats, the relatively low number of exercises, enhanced need for security of the highly memorable content, and the inclusion of highly subjective scoring rubrics were all qualities of performance assessments which Crocker felt posed challenges to the establishment of test validity. She contended that, due to claims that these types of performance represented authentic assessment, most validity arguments would be based on inspection by content experts. Due to the high stakes nature of certification decisions, she called for recognized standards of professional practice for obtaining validity evidence for performance assessments.

**Structured Interviews.** An alternative approach to the assessment of new teachers, several US jurisdictions, including Houston Independent School District, and Milwaukee Public Schools, have implemented a formalized, structured interview, which could be scored on a standardized (checklist) basis. The premise by which the Urban Teacher Selection Interview (UTSI, Haberman, 1991) was designed is that most jurisdictions avail themselves of the interview process when selecting new teachers to hire, so that if a structured, standardized interview could be designed to assess relevant abilities, it would be a cost-effective and valid measure of teacher ability. Haberman identified a series of functions which, he argued, teachers needed to demonstrate in order to be

effective in their positions. The seven functions that were incorporated into the interview were: persistence; response to authority; application of generalizations; approach to at-risk students; personal vs. professional orientation to teaching; susceptibility to burnout; and, recognition of personal fallibility.

Several authors have challenged the predictive validity of this instrument, however, and their challenges appear to stand not only for this particular structured interview instrument, but also for all similar attempts at evaluation. Baskin, Ross, and Smith (1996) assessed the predictive validity of the UTSI for three different groups of new teacher applicants, and reported that only a limited number of significant predictors resulted. These significant findings indicated that those teachers with good communication skills were more likely to experience success as teachers.

Pecheone & Carey (1990) warned that the term "interview" could be misleading, as most people will associate this with the informal type of employment interviews conducted in industry. They contended that a structured interview was more like an oral examination used for graduate degrees, not like the interview typically used by business.

The state of Connecticut developed a semi-structured interview as a sole evaluation tool for beginning teachers, but subsequently decided to expand the instrument to include additional forms of assessment (Pecheone & Carey, 1990). This project was named the Connecticut Teacher Assessment Center Project (CONNTAC), and facilitated the inclusion of portfolios and performance assessment tasks such as structuring a lesson or a unit, choosing an approach to teach a given content, and evaluating exemplars of student performance.

As was the warning from several authors regarding the sole use of paper-and-pencil tests to evaluate teacher effectiveness, educational researchers have echoed this caution for the use of authentic assessment tools. Once again, the reliance on a sole evaluation tool may present serious challenges to the validity and reliability of the certification exercise. With this warning in mind, authentic assessment instruments, along with structured classroom observations, do offer the potential to significantly improve the validity of teacher certification decisions by taking into account teachers' actual on-the-job performance. It should be noted that few of these methods (perhaps only some of the simulations) would be referred to as "tests," but rather as "evaluation methods," so that the term "teacher testing" may unnecessarily limit the (public's, government's , and regulatory bodies') conceptualization of what tasks could contribute to certification decisions. Politically, however, once the public is told that teacher licensure decisions would be made by some method other than the administration of a standardized test, they may feel that teachers have somehow managed to dodge the bullet and continue practicing despite lack of basic teaching competencies. Realistically this would be unlikely if well validated authentic assessment instruments were employed in conjunction with objective classroom observations.

## Conclusion

The above discussion of different methods of assessing teacher competency should lead the reader to conclude that current standardized teacher tests leave much to be desired when it comes to real evidence of psychometric validity. Whether or not the tests could be considered politically valid, however, can be answered primarily by whether the concerns of politicians and the public stay focused on the teaching profession, or turn elsewhere, indicating that the political agenda has been satisfied.

In Massachusetts, for example, the Education Commissioner, in response to the improving test pass rates (being created by educational institutions' pre-testing of teacher education applicants, and by minority applicants' increased reluctance to apply) was quoted as saying

> I continue to be encouraged by the steady improvement of scores, and I am delighted by the increasing numbers of test-takers. This suggests to me that a greater number of qualified candidates are seeking to enter teaching careers. Also notable is the improvement in the pass rate for writing, which I believe reflects the improved preparation of candidates that is taking place in our institutions of higher education (Fowler, 2001, p. 779)

Fowler (2001) challenged this confidence that the test had accomplished so much, by pointing out that the students who were writing the test and achieving the higher pass rates were already in their education programs when the test was initially instituted. In other words, the pass rate increases could not have been due to the tests' causing changes in admission policies because those achieving the higher pass rates were already admitted prior to the test's institution. As well, the students were too far along in their programs for their institutions to have changed their courses substantially in response to the test. Fowler thereby contended that the increase in pass rates was actually artificially achieved by the continued annual lowering of the pass/fail cut-score by the government, and was an indication of how initial pass rates were severely depressed initially in order to garner the inevitable public outrage which ensued when the public was informed that 60% of teacher candidates "failed" the test. It would then be in the political interest of the test administrators (and totally invalid from any psychometric perspective) to systematically lower the pass mark, and create the *appearance* of increasing competence on the parts of teacher candidates--an increase in perceived competence that could then be attributed directly to the existence of the test, affording the test the "political validity" necessary for it to enjoy public and government support. This is in spite of the fact that mounting research into the psychometric validity of the MECT indicates that the test is not actually delivering better teachers (Flippo & Riccards, 2000; Fowler, 2001; Melnick & Pullin, 2000).

Another example of how the establishment pass-fail cut-scores can embroil a test in political controversy is that of an Alabama lawsuit which saw a teacher suing a Board of Education for non-renewal of teaching contract because she did not pass the Alabama Initial Teacher Certification Tests. Her suit was successful because the judge determined that the cut-scores were so poorly justified that they could "only be characterized as capricious and arbitrary." (Horn, Ramos, Blumer & Madaus, 2000, p.10). These cut scores, similar to those of the MECT, were not found to have been established through any psychometrically accepted methods (Jaeger, 1986; Rudner & Eissenberg, 1990). The Alabama cuts scores were, indeed, found to be outrageously high when initially set for the examination, so much so that of the 500 teachers who took the first (standardization) administration of the core examination, none would have passed. When the test developer approached the state with this problem, suggesting that there were several psychometrically complex methods of adjusting and validating the cut-scores, the state chose to merely lower the cut score to create a "politically" acceptable pass rate. This concern for political expedience led to the pass rate depending strictly on an arbitrary number which was not reflective of any psychometrically justifiable concrete determination of minimal competence on the part of test-takers.

Concerns for *ecological* validity (Campbell & Stanley, 1966; Morgan & Gliner, 1997) would also suggest that general academic ability and test-wiseness are often rewarded by paper and pencil tests in testing situations which do not effectively simulate a realistic teaching environment, and thereby do not assess competencies relevant to day-to-day teaching duties. Writing (even the most well developed and reliable) paper-and-pencil test does not allow the licensing body to observe the teacher's effectiveness when actually performing the duties for which they are being certified. Perhaps one of the biggest requirements of a teaching professional is the ability to stand in front of a group and effectively relay often complex material, while constantly performing monitoring and classroom management activities. The isolation, silence, and structure (and relative peace) of the formal written examination could not be farther removed from the reality of a teacher working in a classroom, rendering the assessment ecologically invalid.

These concerns for ecological validity have led many other licensing bodies to design tests which simulate professionals actually performing their duties. Whether this is law candidates researching and writing legal briefs, or medical candidates examining standardized patients during the simulations which comprise their Objective Structured Clinical Examinations (Newble & Swanson, 1988; Resnick, et al., 1992), these professionals are demonstrating the ability to apply their knowledge and skills in relevant professional settings. Concerned parents should be more interested in how a prospective teacher would diffuse a potentially violent situation in a classroom rather than how they would report on an written test what they think the examiner will expect.

The preceding is an example of the differences between professional and public demands for valid testing as described by Melnick and Pullin (2001). While the public may (perhaps quite self-righteously) feel that they are reasonably demanding only the highest of standards their children, and they will be assured of these standards by the government agencies who oversee the quick implementation of these tests, they may actually only be assuring a poor to mediocre educational system. This is due to their not allowing the self-regulation of the teaching profession which would, no doubt, employ far more rigorous, content-valid, criterion-valid and ecologically valid testing methods than those that are quick, easy, and economical, thereby meeting most political agendas. It appears that perhaps many politicians, and the (albeit unwitting) public are seeking quick assurance that everything in the educational system is "alright" so that they can turn their concerns elsewhere, meeting, at least, the standards for political validity of their tests.

## ***REFERENCES***

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington: AERA.

Aksamit, D. & Kluender, M. (1986). A National Perspective of Impact of Basic Skills Testing on Students and Programs. *Teacher Education and Practice, 3*, 33-39.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, D. C.: American Council on Education.

Anrig, G.R. (1987). Teacher Testing in American Education: Useful but no Shortcut to Excellence. *What is the Appropriate Role of Testing in the Teaching Profession? Proceedings of a Cooperative Conference*. Washington, D.C.: National Education Association.

Ayers, J.B. (1988). Another look at the concurrent and predictive validity of the National Teacher Examination. *Journal of Educational Research, 81*, 133-137.

Barton, J., & Collins, A. (1993). Portfolios in teacher education. *Journal of Teacher Education, 44*, 200-210.

Baskin, M., Ross, S, and Smith, X. (1996). Selecting Successful Teachers: The Predictive Validity of the Urban Teacher Selection Interview. *Teacher Educator, 32*, 1-12.

Bird, T. (1990). The schoolteacher's portfolio: An essay of possibilities. In J Millman and L Darling-Hammond (Eds.), *The New Handbook of Teacher Evaluation: Assessing Elementary and Secondary School Teachers, 2nd ed.*, Newbury Pack, CA: SAGE Publications.

Brookhart, S. and Loadman, W. (1992). Teacher Assessment and Validity: What do we want to know? *Journal of Personnel Evaluation in Education, 5*, 347-357.

Campbell, D., & Stanley, J. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Cole, D. J. (1990). *The developing professional: Process and product portfolios*. Paper presented at the annual meeting of the American Association of Colleges of Teacher Education (AACTE), San Antonio, TX.

Conklin, R. C. (1985). Teacher Competency Testing – the Present Situation and Some Concerns on How Teachers are Tested. *Education Canada, Spring*, 12-15.

Crocker, L. (1997). Assessing Content Representativeness of Performance Assessment Exercises. *Applied Measurement in Education, 10*. 83-95.

Cruickshank, D. R., & Metcalf, K.K. (1993). Improving preservice teacher assessment through on-campus laboratory experiences. *Theory into Practice, 32*, 86-92.

Darling-Hammond, L. (1988). The Futures of Teaching. *Educational Leadership, 46*, 11-15, 17.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8* (1).

Darling-Hammond, L., Berry, B. & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*, 55-77.

D'Costa, A. G. (1983). The impact of courts on teacher competence testing. *Theory into Practice, 32*, 104-112.

Dorbry, A.M., Murphy, P.D. & Schmidt, D.M. (1985). Predicting teacher competence. *Action in Teacher Education, 7*, 69-74.

Dybdahl, C., Shaw, D. and Edwards, D. (1997). Teacher Testing: Reason or Rhetoric. *Journal of Research and Development in Education, 30*, 248-254.

Egan, P.J. & Ferre, V.A. (1989). Predicting Performance on the National Teacher Examinations Core Battery. *Journal of Educational Research, 82*, 227-230.

Ellett, C., Loup, K. and Chauvin, S. (1990). *The System for Teaching and learning Assessment and Review (STAR): Annotated guide to teaching and learning*. Baton Rouge: Louisiana State University, College of Education, Louisiana Teaching Internship and Statewide Teacher Evaluation Programs.

Ellett, C., Loup, K. & Chauvin, S. (1991). The Effects of "High Stakes" Certification Demands on the Generalizability and Dependability of a Classroom-based Teacher Assessment System. *Journal of Classroom Interaction, 26*, 25-36.

Fisher, T. H., Fry, B. V., Loewe, K. L., and Wilson, G. W. (1985). Testing teachers for merit pay purposes in Florida. *Educational Measurement: Issues and Practice, 4*, 10-12.

Flippo, R. F. & Riccards, M. P. (2000). Initial teacher certification testing in Massachusetts: A case of the tail wagging the dog. *Phi Delta Kappan, 82*, 34-37.

Fowler, R. C. (2001). What did the Massachusetts Teacher Tests say about American education? *Phi Delta Kappan, 82*, 773-780.

Garcia, I. & Garcia, P.A. (1989). Testing for Failure: Why Children Fail in School and Life. Teacher *Education Quarterly, 16*, 85-91.

Gellman, E. (1993). The Use of Portfolios in Assessing Teacher Competence: Measurement Issues. *Action in Teacher Education, 14*, 39-44.

Haberman, M. (1992). *A brief review of the history and development of the Urban Teacher Selection Interview*. Unpublished manuscript, University of Wisconsin.

Haney, W., Madaus, G. & Kreitzer, A. (1987). Charms talismanic: Testing teachers for the improvement of American Education. In E.Z. Rothkopf (Ed.), *Review of research in education*. Washington, D.C.: American Educational Research Association.

Horn, C., Ramos, I., Blumer, R. & Madaus, G.(2000). Cut scores: Results may vary. *The National Board on Educational Testing and Public Policy Monographs, 1*, 1-31.



BEST COPY AVAILABLE

Hyman, R.R. (1984). Testing for Teacher Competence: The Logic, The Law, and The Implications. *Journal of Teacher Education, 35*, 14-18.

Jaeger, R. M. (1986). Policy issues in standard setting for professional licensure tests. In W. P. Gorth & M. L. Chernoff (Eds.), *Testing for teacher certification*. Hillsdale, NJ: Lawrence Erlbaum.

Klass, D. (1994). "High-Stakes" testing of medical students using standardized patients. *Teaching and Learning in Medicine, 6*, 28-32.

Klein, S. P. (1998). Standards for Teacher Tests. *Journal of Personnel Evaluation in Education, 12*. 123-38.

Klein, S. P. & Stecher, B. (1991). Developing a prototype licensing examination for secondary school teachers. *Journal of Personnel Evaluation in Education, 5*. 169-190.

Loadman, W.E. & Brookhart, S.M. (1987). *Technical Report #6: 1986-1987 results of the NTE*. Columbus: College of Education, The Ohio State University.

Loadman, W.E. & Brookhart, S.M. (1988). *Technical Report #7: 1987-1988 results of the NTE*. Columbus: College of Education, The Ohio State University.

Madaus, G.F. & Kellaghan, T. (1993). The British experience with "authentic" testing. *Phi Delta Kappan, 74*, 458-469.

Madaus, G.F. & Pullin, D. (1987). Teacher Certification tests: Do they really measure what we need to know? *Phi Delta Kappan, 69*, 31-38.

Matthews, P., Holmes, J.R., Vickers, P. & Corporaal, B. (1998). Aspects of the Reliability and Validity of School Inspection Judgements of Teaching Quality. *Educational Research and Evaluation, 4*, 167-188.

McConney, A.A., Schalock, M.D. & Schalock, H.D. (1998). Focusing Improvement and Quality Assurance: Work Samples as Authentic Performance Measures of Prospective Teachers' Effectiveness. *Journal of Personnel Evaluation in Education, 11*, 343-363.

Melnick, S. L. & Pullin, D. (2000). Can you take dictation? Prescribing teacher quality through testing. *Journal of Teacher Education, 51*, 262-275.

Melnick, S. L. (1996). Reforming teacher education through legislation: A case study from Florida. In K. Zeichner, S. Melnick, & M. L. Gomes (Eds.) *Currents of reform in preservice teacher education* (30-61). New York: Teachers College Press.

Moore, D. (1991). Correlations of National Teacher Examination Core Battery scores and College Grade Point Average with Teaching Effectiveness of First-year Teachers. Educational and Psychological Measurement, Inc. 51.

Moore, D., Schurr, K.T. & Henriksen, L.W. (1991). Correlations of the National Teacher Examination Core Battery Scores and College Grade Point Average with Teaching Effectiveness of First-Year Teachers. *Educational and Psychological Measurement, 51*, 1023-1028.

Morgan, G.A. & Gliner, J.A. (1997). *Helping Students Evaluate the Validity of a Research Study*. Paper presented at the Annual Meeting of the AERA, Chicago, Il, March 24-28, 1997).

Newble, D.I. & Swanson, D.B. (1988). Psychometric characteristics of the OSCE. *Medical Education, 22*, 325-334.

Olstad, R.G. (1983). *Preservice teaching performance: A search for predictor variables. Final Report*. (Report No. 83-3). Seattle, WA: Washington University, Teacher Education Research Center. (ERIC Document Reproduction Service No. ED 231 810).

Olstad, R.G., Beal, J.L., & Marrett, A.V. (1988). *The relationship of NTE exams to teacher education admission, performance, and employment*. Seattle: College of Education, University of Washington.

Pecheone, R.L & Carey, N.B. (1990). The Validity of Performance Assessments for Teacher Licensure: Connecticut's Ongoing Research. *Journal of Personnel Evaluation in Education, 3*, 115-142.

Piper, M.K. & Houston, R.W. (1980). The Search for Teacher Competence: CBTE and MCT. *Journal for Teacher Education, 31*, 37-40.

Poggio, J., Glasnapp, D.R. Green, S.B., & Tollefson, N. (1997). *Conducting Licensure Validity Studies: The Need to Broaden the Evidentiary Base*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

Pratt, L.K., DeLucia, S. & Williams, V.S.L. (1987). *Predicting student performance on the Professional Knowledge portion of the NTE Core Battery*. Paper presented at the Annual Conference of the Association of Institutional Research, New Orleans, LA.

Quirk, T., Witten, B., & Weinberg, S. (1973). Review studies of the concurrent and predictive validity of the NTE. *Review of Educational Research, 43*, 89-113.

Reznick, R., Smee, S., Rothman, A. et al. (1992). An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Academic Medicine, 67*, 487-494.

Roberson, T.J. (1998). *Classroom Observation: Issues Regarding Validity and Reliability*. Paper presented at the Annual Meeting of the Mid-South Education Research Association, New Orleans, LA.

Rudner, L. (1988). Teacher Testing – An Update. *Educational Measurement: Issues and Practice*, 7, 16-19.

Rudner, L.M. & Eissenberg, T. E. (1990). Standard-setting practices for teacher tests. *Journal of Personnel Evaluation in Education, 3*, 143-149.

Ryan, J.M., & Kuh, T.M. (1993). Assessment of preservice teachers and the use of portfolios. *Theory into Practice, 32*, 75-81.

Scriven, M. (1988). Duty-based evaluation. Journal of Personnel Evaluation in Education, 1, 319-334.

Shannon, D.M. and Boll, M. (1996). Assessment of pre-service teachers using alternative assessment methods. *Journal of Personnel Evaluation in Education, 10*, 117-135.

Shepard, L.A. (1989). Why we need better assessment. *Educational Leadership, 46*, 2-9.

Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4-14.

Shulman, L.S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1-22.

Shulman, L.S. (1989). The Paradox of Teacher Assessment. *New Directions for Teacher Assessment* (13-27). Princeton, N.J.: Educational Testing Service.

Stivers J. & McMorris, R. F. (1991). Relating a Test for Teachers to Research Literature on Teaching Effectiveness. *Journal of Personnel Evaluation in Education, 5*, 31-53.

Violato, C. McDougall, D. & Marini, A. (1992). Educational Measurement and Evaluation. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Wakeford, M.E. (1988). The incremental predictive validity of parts I and II of the National Teacher Examinations (NTE) Core Battery. (Doctoral dissertation, University of North Carolina at Chapel Hill, 1987). *Dissertation Abstracts International, 49*, 2613A.



BEST COPY AVAILABLE

Worthen, B.R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan, 74,* 444-454.

Zimpher, W., Chair Task Force on Certification Examinations. (1990*). A Report from the Task Force on Certification Examinations.* Columbus, OH: The Ohio State University, College of Education.